

BAYESIAN STATISTICS 9,

J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,

D. Heckerman, A. F. M. Smith and M. West (Eds.)

© Oxford University Press, 2010

Free energy Sequential Monte Carlo, application to mixture modelling

N. CHOPIN & P. JACOB

CREST (ENSAE), France

nicolas.chopin@ensae.fr, pierre.jacob@ensae.fr

SUMMARY

We introduce a new class of Sequential Monte Carlo (SMC) methods, which we call free energy SMC. This class is inspired by free energy methods, which originate from Physics, and where one samples from a biased distribution such that a given function $\xi(\theta)$ of the state θ is forced to be uniformly distributed over a given interval. From an initial sequence of distributions (π_t) of interest, and a particular choice of $\xi(\theta)$, a free energy SMC sampler computes sequentially a sequence of biased distributions $(\tilde{\pi}_t)$ with the following properties: (a) the marginal distribution of $\xi(\theta)$ with respect to $\tilde{\pi}_t$ is approximatively uniform over a specified interval, and (b) $\tilde{\pi}_t$ and π_t have the same conditional distribution with respect to ξ . We apply our methodology to mixture posterior distributions, which are highly multimodal. In the mixture context, forcing certain hyper-parameters to higher values greatly facilitates mode swapping, and makes it possible to recover a symmetric output. We illustrate our approach with univariate and bivariate Gaussian mixtures and two real-world datasets.

Keywords and Phrases: FREE ENERGY BIASING; LABEL SWITCHING; MIXTURE; SEQUENTIAL MONTE CARLO; PARTICLE FILTER.

1. INTRODUCTION

A Sequential Monte Carlo (SMC) algorithm (a.k.a. particle filter) samples iteratively a sequence of probability distributions $(\pi_t)_{t=0,\dots,T}$, through importance sampling and resampling steps. The initial motivation of SMC was the sequential analysis of dynamic state space models, where π_t stands for the filtering distribution of state (latent variable) x_t , conditional on the data $y_{1:t}$ collected up to time t ; see e.g. the book of Doucet et al. (2001). Recent research however (Neal, 2001; Chopin, 2002; Del Moral et al., 2006) have extended SMC to “static” problems, which involves a single, but “difficult” (in some sense we detail below) distribution π . Such extensions use an artificial sequence $(\pi_t)_{t=0,\dots,T}$, starting at some “simple” distribution π_0 , and evolving smoothly towards $\pi_T = \pi$. Two instances of such strategies are i) annealing (Neal 2001, see also Gelman and Meng 1998), where $\pi_t(\theta) = \pi_0(\theta)^{1-\gamma_t} \pi(\theta)^{\gamma_t}$, and $\gamma_t = t/T$, or some other increasing sequence that starts at 0 and ends at 1; and ii) IBIS (Chopin, 2002), where π stands for some Bayesian posterior density $\pi(\theta) = p(\theta|y_{1:T})$, conditional on some complete dataset $y_{1:T}$, and $\pi_t(\theta) = p(\theta|y_{1:t})$. For a general formalism for SMC, see Del Moral et al. (2006).

One typical “difficulty” with distributions of interest π is multimodality. A vanilla sampler typically converges to a single modal region, and fails to detect other modes, which may be of higher density. The two SMC strategies mentioned above alleviate this problem to some extent. In both cases, π_0 is usually unimodal and has a large

Support from the ANR grant ANR-008-BLAN-0218 by the French Ministry of research is acknowledged.

support, so “particles” (sampled points) explore the sample space freely during the first iterations. However, this initial exploration is not always sufficient to prevent the sample to degenerate to a single modal region. We give an illustration of this point in this paper.

To overcome multimodality, the molecular dynamics community has developed in recent years an interesting class of methods, based on the concept of free energy biasing; see for instance the book of Lelièvre et al. (2010) for a general introduction. Such methods assume the knowledge of a low-dimensional function $\xi(\theta)$, termed as the “reaction coordinate”, such that, conditional on $\xi(\theta) = x$, the multimodality (a.k.a. metastability in the physics literature) of π is much less severe, at least for certain values of x . The principle is then to sample from $\tilde{\pi}$, a free energy biased version of π , $\tilde{\pi}(\theta) = \pi(\theta) \exp \{A \circ \xi(\theta)\}$, where A denotes the free energy, that is, minus log the marginal density of the random variable $\xi(\theta)$, with respect to π . This forces an uniform exploration of the random variable $\xi(\theta)$, within given bounds. At a final stage, one may perform importance sampling from $\tilde{\pi}$ to π to recover the true distribution π .

The main difficulty in free energy biasing methods is to estimate the free energy A . A typical approach is to compute sequentially an estimate $A_{(t)}$ of A , using some form of Adaptive MCMC (Markov chain Monte Carlo): at each iteration t , a MCMC step is performed, which leaves invariant $\pi_{(t)}(\theta) = \pi(\theta) \exp \{A_{(t)} \circ \xi(\theta)\}$, then a new estimate $A_{(t+1)}$ of the free energy is computed from the simulated process up to time t . The simulation is stopped when the estimate $A_{(t)}$ stabilises in some sense. Convergence of Adaptive MCMC samplers is a delicate subject: trying to learn too quickly from the past may prevent convergence for instance. These considerations are outside the scope of this paper, and we refer the interested reader to the review by Andrieu and Thoms (2008) and references therein.

Instead, our objective is to bring the concept of free energy biasing to the realm of SMC. Specifically, and starting from some pre-specified sequence (π_t) , we design a class of SMC samplers, which compute sequentially the free energy A_t associated to each distribution π_t , and track the sequence of biased densities $\tilde{\pi}_t(\theta) = \pi_t(\theta) \exp \{A_t \circ \xi(\theta)\}$. In this way, particles may move freely between the modal regions not only at the early iterations, where π_t remains close to π_0 and therefore is not strongly multimodal, but also at the later stages, thanks to free energy biasing.

We apply free energy SMC sampling to the Bayesian analysis of mixture models. Chopin et al. (2010) show that free energy biasing methods are an interesting approach for dealing with the multimodality of mixture posterior distributions. In particular, they present several efficient reaction coordinates for univariate Gaussian mixtures, such as the hyper-parameter that determines the prior expectation of the component variances. In this paper, we investigate how free energy SMC compares with this initial approach based on Adaptive MCMC, and to which extent such ideas may be extended to other mixture models, such as a bivariate Gaussian mixture model.

The paper is organised as follows. Section 2 describes the SMC methodology. Section 3 presents the concept of free energy biased sampling. Section 4 presents a new class of SMC methods, termed as free energy SMC. Section 5 discusses the application to Bayesian inference for mixtures, and presents numerical results, for two types of mixtures (univariate Gaussian, bivariate Gaussian), and two datasets. Section 6 concludes.

2. SMC ALGORITHMS

2.1. Basic structure

In this section, we describe briefly the structure of SMC algorithms. For the sake of exposition, we consider a sequence of probability densities π_t , $t = 0, \dots, T$ defined on a common space $\Theta \subset \mathbb{R}^d$. At each iteration t , a SMC algorithm produces a weighted sample $(w_{t,n}, \theta_{t,n})$, $n = 1, \dots, N$, which targets π_t in the following sense:

$$\frac{\sum_{n=1}^N w_{t,n} \varphi(\theta_{t,n})}{\sum_{n=1}^N w_{t,n}} \rightarrow_{N \rightarrow +\infty} \mathbb{E}^{\pi_t} \{\varphi(\theta)\},$$

almost surely, for a certain class of test functions φ . At iteration 0, one typically samples $\theta_{0,n} \sim \pi_0$, and set $w_{0,n} = 1$. To progress from iteration $t-1$ to iteration t , it

is sufficient to perform a basic importance sampling step from π_{t-1} to π_t :

$$\theta_{t,n} = \theta_{t-1,n}, \quad w_{t,n} = w_{t-1,n} \times u_t(\theta_{t,n})$$

where u_t denotes the incremental weight function

$$u_t(\theta) = \frac{\pi_t(\theta)}{\pi_{t-1}(\theta)}.$$

However, if only importance sampling steps are performed, the algorithm is equivalent to a single importance sampling step, from π_0 to π_T . This is likely to be very inefficient. Instead, one should perform regularly resample-move steps (Gilks and Berzuini, 2001), that is, a succession of i) a resampling step, where current points $\theta_{t,n}$ are resampled according to their weights, so that points with a small (resp. big) weight are likely to die (resp. generate many offsprings); and ii) a mutation step, where each resampled point is “mutated” according to some probability kernel $K_t(\theta, d\hat{\theta})$, typically a MCMC kernel with invariant distribution π_t . In the more general formalism of Del Moral et al. (2006), this is equivalent to performing an importance sampling step in the space $\Theta \times \Theta$, with forward kernel K_t , associated to some probability density $K_t(\theta, \hat{\theta})$, and backward kernel L_t associated to the probability density $L_t(\hat{\theta}, \theta) = \pi_t(\theta)K_t(\theta, \hat{\theta})/\pi_t(\hat{\theta})$.

Resample-move steps should be performed whenever the weight degeneracy is too high. A popular criterion is $\text{EF}(t) < \tau$, where $\tau \in (0, 1)$, and EF is the efficiency factor, that is the effective sample size of Kong et al. (1994) divided by N ,

$$\text{EF}(t) = \frac{\left(\sum_{n=1}^N w_{t,n}\right)^2}{N \sum_{n=1}^N w_{t,n}^2}.$$

We summarise in Algorithm 1 the general structure of SMC algorithms. There are several methods for resampling the particles, e.g. the multinomial scheme (Gordon et al., 1993), the residual scheme (Liu and Chen, 1998), the systematic scheme (Whitley, 1994; Carpenter et al., 1999). We shall use the systematic scheme in our simulations.

2.2. Adaptiveness of SMC

In contrast to MCMC, where designing adaptive algorithms require a careful convergence study, designing adaptive SMC samplers is straightforward. We consider first the design of the MCMC kernels K_t . For instance, Chopin (2002) uses independent Hastings-Metropolis kernels, with a Gaussian proposal fitted to the current particle sample. This is a reasonable strategy if π_t is close to Gaussianity. In this paper, we consider instead the following strategy, which seems more generally applicable: take K_t as a succession of k Gaussian random walk Hastings-Metropolis steps $K_{t,i}(\theta, d\theta')$, i.e. simulating from $K_{t,i}(\theta, d\theta')$ consists of proposing a value $\theta' \sim N_d(\theta, \Sigma_{t,i})$, accepting this value with probability $1 \wedge \{\pi(\theta')/\pi(\theta)\}$, otherwise keep the current value θ . Then take $\Sigma_{t,i} = c_{t,i}S_t$, $c_{t,i} > 0$, and S_t is the empirical covariance matrix of the resampled particles at iteration t (that is, the particles obtained immediately before the MCMC step with kernel K_t is performed). The constant $c_{t,i}$ may be tuned automatically as well. For instance, one may start with $c_0 = 0.3$, then, each time the acceptance rate of the MCMC step is below (resp. above) a given threshold, the constant c_t is divided (resp. multiplied) by two.

As in MCMC, it is common to focus on the adaptiveness of the transition kernels, but one may use the particle sample (or the history of the process in the MCMC context) to adapt the target distributions as well. This is precisely what we do in this paper, since the target at time t on our free energy SMC sampler shall depend on a bias function which is estimated from the current particle sample, see Section 4.

2.3. IBIS versus annealing, choice of π_0

When the distribution of interest π is some Bayesian posterior density

$$\pi(\theta) = p(\theta|y_{1:D}) = \frac{1}{Z}p(\theta)p(y_{1:D}|\theta),$$

Algorithm 1 A generic SMC algorithm

0. Sample $\theta_{0,n} \sim \pi_0$, set $w_{0,n} = 1$, for $n = 1, \dots, N$. Set $t = 1$.

1. Compute new weights as

$$w_{t,n} = w_{t-1,n} \times u_t(\theta_{t-1,n}).$$

2. If $\text{EF}(t) < \tau$, then

(a) resample the particles, i.e. construct a sample $(\hat{\theta}_{t,n})_{1 \leq n \leq N}$ made of $R_{t,n}$ replicates of particle $\theta_{t,n}$, $1 \leq n \leq N$, where $R_{t,n}$ is a nonnegative integer-valued random variable such that

$$\mathbb{E}[R_{t,n}] = \frac{N w_{t,n}}{\sum_{n'=1}^N w_{t,n'}},$$

and set $w_{t,n} = 1$.

(b) move the particles with respect to Markov kernel K_t ,

$$\theta_{t,n} \sim K_t(\hat{\theta}_{t,n}, d\theta)$$

otherwise

$$\theta_{t,n} = \theta_{t-1,n}.$$

3. $t \leftarrow t + 1$, if $t < T$ go to Step 1.

where $y_{1:D}$ is a vector of D observations, $p(\theta)$ is the prior density, and $p(y_{1:D}|\theta)$ is the likelihood, it is of interest to compare the two aforementioned SMC strategy, namely,

1. IBIS, where $T = D$, and $\pi_t(\theta) = p(\theta|y_{1:t})$, in particular, $\pi_0(\theta) = p(\theta)$ is the prior; and
2. Annealing, where $\pi_t(\theta) = \pi_0(\theta)^{1-\gamma_t} \pi(\theta)^{\gamma_t}$, γ_t is an increasing sequence such that $\gamma_0 = 0$, and $\gamma_T = 1$, π_0 is typically the prior density, but could be something else, and T and D do not need to be related.

Clearly, for the same number of particles, and assuming that the same number of resample-move steps is performed, IBIS is less time-consuming, because calculations at iteration t involve only the t first observations. On the other hand, annealing may produce a smoother sequence of distributions, so it may require less resample-move steps. Jasra et al. (2007) provide numerical examples where the IBIS strategy leads to unstable estimates. In the context discussed in the paper, see Section 5, and elsewhere, we did not run into cases where IBIS is particularly unstable. Perhaps it is fair to say that a general comparison is not meaningful, as the performance of both strategies seems quite dependent on the applications, and also various tuning parameters such as the sequence γ_t for instance.

We take this opportunity however to propose a simple method to improve the regularity of the IBIS sequence, in the specific case where the observations are exchangeable and real-valued. We remark first that this regularity depends strongly on the order of incorporation of the y_t 's. For instance, sorting the observations in ascending order would certainly lead to very poor performance. On the other hand, a random order would be more suitable, and was recommended by Chopin (2002). Pushing this idea further, we propose the following strategy: First, we re-define the median of a sample as either the usual median, when D is an odd number, or the smallest of the two middle values in the ordered sample, when D is an even number. Then, we take y_1 as the median observation, y_2 (resp. y_3) to be the median of the observations that are smaller (resp. larger) than y_1 , then we split again the four corresponding sub-samples by selecting some values y_4 to y_7 , and so on, until all values are selected. We term

this strategy as “Van der Corput ordering”, as a Van der Corput binary sequence is precisely defined as $1/2, 1/4, 3/4, 1/8, \dots$

A point which is often overlooked in the literature, and which affects both strategies, is the choice of π_0 . Clearly, if $\pi_0(\theta) = p(\theta)$, one may take the prior so uninformative that the algorithm degenerates in one step. Fortunately, in the application we discuss in this paper, namely Bayesian analysis of mixture models, priors are often informative; see Section 5 for a discussion of this point. In other contexts, it may be helpful to perform a preliminary exploration of π in order design some π_0 , quite possibly different from the prior, so that (i) for the annealing strategy, moving from π_0 to $\pi_T = \pi$ does not take too much time; and (ii) for the IBIS strategy, one can use π_0 as an artificial prior, and recover the prior of interest at the final stage of the algorithm, by multiplying all the particle weights by $p(\theta)/\pi_0(\theta)$.

3. FREE ENERGY-BIASED SAMPLING

3.1. Definition of free energy, and free-energy biased densities

In this section we explain in more detail the concept of free energy biasing. We consider a single distribution of interest, defined by a probability density π with respect to the Lebesgue measure associated to $\Theta \subset \mathbb{R}^d$. As explained in the introduction, the first step in implementing a free energy biasing method is to choose a reaction coordinate, that is, some measurable function $\xi : \theta \rightarrow \mathbb{R}^{d'}$, where d' is small. In this paper, we take $d' = 1$. One assumes that the multimodality of π is strongly determined, in some sense, by the direction $\xi(\theta)$. For instance, the distribution of θ , conditional on $\xi(\theta) = x$, may be much less multimodal than the complete distribution π , for either all or certain values of x .

In words, the free energy is, up to an arbitrary constant, minus the logarithm of the marginal density of $\xi(\theta)$. The free energy may be written down informally as

$$\exp \{-A(x)\} \propto \int_{\Theta} \pi(\theta) \mathbf{I}_{[x, x+dx]} \{\xi(\theta)\} d\theta$$

and more rigorously, as

$$\exp \{-A(x)\} \propto \int_{\Omega_x} \pi(\theta) d\{\theta | \xi(\theta) = x\},$$

where $\Omega_x = \{\theta \in \Theta : \xi(\theta) = x\}$, and $d\{\theta | \xi(\theta) = x\}$ denotes the conditional measure on the set Ω_x which is “compatible” with Lebesgue measure on the embedding space Θ , i.e. volumes are preserved and so on. In both formulations, the proportionality relation indicates that the density π may be known only up to a multiplicative constant, and therefore that the free energy is defined only up to an arbitrary additive constant.

The free energy biased density $\tilde{\pi}$ is usually defined as

$$\tilde{\pi}(\theta) \propto \pi(\theta) \exp \{A \circ \xi(\theta)\} \mathbf{I}_{[x_{\min}, x_{\max}]} \{\xi(\theta)\}$$

where $[x_{\min}, x_{\max}]$ is some pre-defined range. It is clear that, with respect to $\tilde{\pi}$, the marginal distribution of the random variable $\xi(\theta)$ is uniform over $[x_{\min}, x_{\max}]$, and the conditional distributions of θ , given $\xi(\theta) = x$ matches the same conditional distribution corresponding to $\tilde{\pi}$. The objective is to sample from $\tilde{\pi}$, which requires to estimate the free energy A .

To avoid the truncation incurred by the restriction to interval $[x_{\min}, x_{\max}]$, we shall consider instead the following version of the free-energy biased density $\tilde{\pi}_t$:

$$\tilde{\pi}_t(\theta) \propto \pi(\theta) \exp \{A \circ \xi(\theta)\}$$

where the definition of A is extended as follows: $A(x) = A(x_{\min})$ for $x \leq x_{\min}$, $A(x) = A(x_{\max})$ for $x \geq x_{\max}$.

3.2. Estimation of the free energy

As explained in the introduction, one usually resorts to some form of Adaptive MCMC to estimate the free energy A . Specifically, one performs successive MCMC steps (typically Hastings-Metropolis), such that the Markov kernel $K_{(t)}$ used at iteration t depends on the trajectory of the simulated process up to time $t - 1$. (The simulated process is therefore non-Markovian.) The invariant distribution of kernel $K_{(t)}$ is $\pi_{(t)}(\theta) \propto \pi(\theta) \exp \{A_{(t)} \circ \xi(\theta)\}$, where $A_{(t)}$ is an estimate of the free energy A that has been computed at iteration t , from the simulated trajectory up to time $t - 1$. Note that the brackets in the notations $K_{(t)}$, $\pi_{(t)}$, $A_{(t)}$ indicate that all these quantities are specific to this section and to the Adaptive MCMC context, and must not mistaken for the similar quantities found elsewhere in the paper, such as, e.g. the density π_t targeted at iteration t by a SMC sampler. The difficulty is then to come up with an efficient estimator (or rather a sequence of estimators, $A_{(t)}$), of the free energy.

Since this paper is not concerned with adaptive MCMC, we consider instead the much simpler problem of estimating the free energy A from a weighted sample $(\theta_n, w_n)_{n=1, \dots, N}$ targeting π ; for instance, the θ_n 's could be i.i.d. with probability density g , and $w_n = \pi(\theta)/g(\theta)$. Of course, this discussion is simplistic from an Adaptive MCMC perspective, but it will be sufficient in our SMC context. We refer the reader to e.g. Chopin et al. (2010) for the missing details.

First, it is necessary to discretise the problem, and consider some partition:

$$[x_{\min}, x_{\max}] = \cup_{i=0}^{n_x} [x_i, x_{i+1}], \quad x_i = x_{\min} + (x_{\max} - x_{\min}) \frac{i}{n_x}. \quad (1)$$

Then, they are basically two ways to estimate A . The first method is to estimate directly a discretised version of A , by simply computing an estimate the proportion of points that fall in each bin:

$$\exp \{-\hat{A}_1(x)\} = \frac{\sum_{n=1}^N w_n \mathbf{I} \{\xi(\theta_n) \in [x_i, x_{i+1}]\}}{\sum_{n=1}^N w_n}, \quad \text{for } x \in [x_i, x_{i+1}].$$

The second method is indirect, and based on the following property: the derivative of the free energy is such that

$$A'(x) = \mathbb{E}^\pi [f(\theta) | \xi(\theta) = x]$$

where the force f is defined as:

$$f = -\frac{(\nabla \log \pi) \cdot (\nabla \xi)}{|\nabla \xi|^2} - \operatorname{div} \left(\frac{\nabla \xi}{|\nabla \xi|^2} \right),$$

and ∇ (resp. div) is the gradient (resp. divergence) operator. Often, $\xi(\theta)$ is simply a coordinate of the vector θ , $\theta = (\xi, \dots)$, in which case the expression above simplifies to $f = -\partial \log \pi / \partial \xi$. This leads to the following estimator of the derivative of A :

$$\hat{A}'_2(x) = \frac{\sum_{n=1}^N w_n \mathbf{I} \{\xi(\theta_n) \in [x_i, x_{i+1}]\} f(\theta_n)}{\sum_{n=1}^N w_n \mathbf{I} \{\xi(\theta_n) \in [x_i, x_{i+1}]\}}, \quad \text{for } x \in [x_i, x_{i+1}].$$

Then an estimate of A may be deduced by simply computing cumulative sums for instance:

$$\hat{A}_2(x) = \sum_{j: x_j \leq x} \hat{A}'_2(x_j)(x_{j+1} - x_j), \quad \text{for } x \in [x_i, x_{i+1}].$$

Methods based on the first type of estimates are usually called ABP (Adaptive Biasing Potential) methods, while methods of the second type are called ABF (Adaptive Biasing Force). Empirical evidence suggests that ABF leads to slightly smoother estimates, presumably because it is based on a derivative.

4. FREE ENERGY SMC

We now return to the SMC context, and consider a pre-specified sequence (π_t) . Our objective is to derive a SMC algorithm which sequentially compute the free energy A_t associated to each density π_t ,

$$\exp \{-A_t(x)\} \propto \int \pi_t(\theta) d\{\theta | \xi(\theta) = x\}$$

and sample $\tilde{\pi}_t$, the free energy biased version of π_t ,

$$\tilde{\pi}_t(\theta) \propto \pi_t(\theta) \exp \{A_t \circ \xi(\theta)\}.$$

Again, to avoid truncating to interval $[x_{\min}, x_{\max}]$, one extends the definition of A_t outside $[x_{\min}, x_{\max}]$ by taking $A_t(x) = A_t(x_{\min})$ for $x < x_{\min}$, $A_t(x) = A_t(x_{\max})$ for $x > x_{\max}$.

As explained in Section 3.2, one actually estimates a discretised version of the free energy, i.e., the algorithm shall provide estimates $\hat{A}_t(x_i)$, $i = 0, \dots, n_x$ of the free energy evaluated at grid points over an interval $[x_{\min}, x_{\max}]$, as defined in (1). Note that this grid is the same for all iterations t .

Assume that we are at the end of iteration $t - 1$, that estimates $\hat{A}_{t-1}(x_i)$ of A_{t-1} have been obtained, and that the particle system $(\theta_{t-1,n}, w_{t-1,n})_{n=1, \dots, N}$ targets $\tilde{\pi}_{t-1}$. If the particles are re-weighted according to the incremental weight function $u_t(\theta) = \pi_t(\theta)/\pi_{t-1}(\theta)$, i.e.

$$\bar{w}_{t,n} = w_{t-1,n} \times u_t(\theta_{t-1,n})$$

then the new target distribution of the particle system $(\theta_{t-1,n}, \bar{w}_{t,n})_{n=1, \dots, N}$ is

$$\tilde{\pi}_t(\theta) \propto \tilde{\pi}_{t-1}(\theta) u_t(\theta).$$

The objective is then to recover $\tilde{\pi}_t$, which depends on the currently unknown free energy A_t . To that effect, we first state the following result.

Theorem 1 *The free energy D_t associated to $\tilde{\pi}_t$ is*

$$D_t = A_t - A_{t-1}$$

that is, the difference between the free energies of π_t and π_{t-1} .

Proof. One has, for $\theta \in \Theta$,

$$\tilde{\pi}_t(\theta) \propto \pi_t(\theta) \exp \{A_{t-1} \circ \xi(\theta)\}$$

hence, for $x \in \xi(\Theta)$,

$$\int_{\Omega_x} \tilde{\pi}_t(\theta) d\{\theta | \xi(\theta) = x\} = \exp \{(A_{t-1} - A_t)(x)\}.$$

and one concludes. \square

This result provides the justification for the following strategy. First, particles are reweighted from π_{t-1} to $\tilde{\pi}_t$, as explained above. Second, the free energy D_t of $\tilde{\pi}_t$ is estimated, using either the ABP or the ABF strategy, see Section 3.2; this leads to some estimate \hat{D}_t of D_t , or more precisely estimates $\hat{D}_t(x_i)$ over the grid x_0, \dots, x_{n_x} . From this, one readily obtains estimates of the current free energy, using the proposition above:

$$\hat{A}_t(x_i) = \hat{A}_{t-1}(x_i) + \hat{D}_t(x_i), \quad i = 0, \dots, n_x. \quad (2)$$

Third, one recovers $\tilde{\pi}_t$ by performing an importance sampling step from $\tilde{\pi}_t$ to $\tilde{\pi}_t$; this is equivalent to updating the weights as follows:

$$w_{t,n} = \bar{w}_{t,n} \exp \left\{ \hat{D}_t \circ \xi(\theta_{t,n}) \right\}.$$

An outline of this free energy SMC algorithm is given in Algorithm 2.

Algorithm 2 Free energy SMC

0. Sample $\theta_{0,n} \sim \pi_0$, set $w_{0,n} = 1$, for $n = 1, \dots, N$. Compute A_0 and set $t = 1$.
 1. Compute new weights as

$$\bar{w}_{t,n} = w_{t-1,n} \times u_t(\theta_{t-1,n}).$$

2. Compute an estimator \hat{D}_t of free energy D_t , compute weights

$$w_{t,n} = \bar{w}_{t,n} \exp \left[\hat{D}_t \circ \xi(\theta_{t-1,n}) \right]$$

and update the estimate \hat{A}_t of the free energy A_t , using (2).

3. If $\text{EF}(t) < \tau$, then

(a) resample the particles, i.e. draw randomly $\hat{\theta}_{t,n}$ in such a way that

$$\mathbb{E} \left[\sum_{n'=1}^N \mathbf{I}(\hat{\theta}_{t,n'} = \theta_{t-1,n}) \mid (\theta_{t-1,n}, w_{t,n}) \right] = \frac{N w_{t,n}}{\sum_{n'=1}^N w_{t,n'}}$$

and set $w_{t,n} = 1$.

- (b) move the particles with respect to Markov kernel K_t ,

$$\theta_{t,n} \sim K_t(\hat{\theta}_{t,n}, d\theta)$$

otherwise

$$\theta_{t,n} = \theta_{t-1,n}.$$

4. $t \leftarrow t + 1$, if $t < T$ go to Step 1.
-

At the final stage of the algorithm (iteration T), one recovers the unbiased target $\pi_T = \pi$ by a direct importance sampling step, from $\tilde{\pi}_T$ to π_T :

$$\frac{\pi_T(\theta)}{\tilde{\pi}_T(\theta)} \propto \exp \left\{ \hat{A}_T \circ \xi(\theta) \right\}.$$

This is because of this ultimate debiasing step, which relies on \hat{A}_T , that one must store in memory and compute iteratively the “complete” free energy A_T (as opposed to the successive D_t , which may be termed as “incremental” free energies). If this unbiasing step is too “brutal”, meaning that too many particles get a low weight in the final sample, then one may apply instead a progressive unbiasing strategy, by extending the sequence of distributions $\tilde{\pi}_T$ as follows:

$$\tilde{\pi}_{T+l}(\theta) \propto \tilde{\pi}_T(\theta) \exp \left\{ \left(\frac{l}{L} \right) \hat{A}_T \circ \xi(\theta) \right\}, \quad l = 0, \dots, L$$

and performing additional SMC steps, that is, successive importance sampling steps from $\tilde{\pi}_{T+l}$ to $\tilde{\pi}_{T+l+1}$, and, when necessary, resample-move steps in order to avoid degeneracy. In our simulations, we found that progressive unbiasing did lead to some improvement, but that often direct unbiasing was sufficient. Hence, we report only results from direct unbiasing in the next Section.

5. APPLICATION TO MIXTURES

5.1. General formulation, multimodality

A K -component Bayesian mixture model consists of D independent and identically distributed observations y_i , with parametric density

$$p(y_i|\theta) = \frac{1}{\sum_{k=1}^K \omega_k} \sum_{k=1}^K \omega_k \psi(y_i; \xi_k), \quad \omega_k \geq 0,$$

where $\{\psi(\cdot; \xi), \xi \in \Xi\}$ is some parametric family, e.g. $\psi(y, \xi) = N(y; \mu, 1/\lambda)$, $\xi = (\mu, \lambda^{-1})$. The parameter vector contains

$$\theta = (\omega_1, \dots, \omega_K, \xi_1, \dots, \xi_K, \eta),$$

where η is the set of hyper-parameters that are shared by the K components. The prior distribution $p(\theta)$ is typically symmetric with respect to component permutation. In particular, one may assume that, a priori and independently $\omega_k \sim \text{Gamma}(\delta, 1)$. This leads to a Dirichlet $_K(\delta, \dots, \delta)$ prior for the component probabilities

$$q_k = \frac{\omega_k}{\sum_{l=1}^K \omega_l}, \quad k = 1, \dots, K.$$

We note in passing that, while the formulation of a mixture model in terms of the q_k 's is more common, we find that the formulation in terms of the unnormalised weights ω_k is both more tractable (because it imposes symmetry in the notations) and more convenient in terms of implementation (e.g. designing Hastings-Metropolis steps).

An important feature of the corresponding posterior density

$$\pi(\theta) = p(\theta|y_{1:D}) \propto p(\theta) \prod_{i=1}^D p(y_i|\theta),$$

assuming D observations are available, is its invariance with respect to “label permutation”. This feature and its bearings to Monte Carlo inference have received a lot of attention, see e.g. Celeux et al. (2000), Jasra et al. (2005), Chopin et al. (2010) among others. In short, a standard MCMC sampler, such as the Gibbs sampler of Diebolt and Robert (1994), see also the book of Frühwirth-Schnatter (2006), typically

visits a single modal region. But, since the posterior is symmetric, any mode admits $K! - 1$ replicates in Θ . Therefore, one can assert that the sampler has not converged. Frühwirth-Schnatter (2001) proposes to permute randomly the components at each iteration. However, Jasra et al. (2005) mentions the risk of “genuine multimodality”, that is, the $K!$ symmetric modal regions visited by the permutation sampler may still represent a small part of the posterior mass, because other sets of equivalent modes have not been visited. (Marin and Robert, 2007, Chap. 6) and Chopin et al. (2010) provide practical examples of this phenomenon.

One could say that random permutations merely “cure the most obvious symptom” of failed convergence. We follow Celeux et al. (2000), Jasra et al. (2005) and Chopin et al. (2010), and take the opposite perspective that one should aim at designing samplers that produce a nearly symmetric output (with respect to label switching), *without resorting to random permutations*.

5.2. Univariate Gaussian mixtures

5.2.1. Prior, reaction coordinates

We first consider a univariate Gaussian mixture model, i.e. $\psi(y, \xi) = N(y; \mu, \lambda^{-1})$, $\xi = (\mu, \lambda^{-1})$, and we use the same prior as in Richardson and Green (1997), that is, for $k = 1, \dots, K$, independently,

$$\mu_k \sim N(M, \kappa^{-1}), \quad \lambda_k \sim \text{Gamma}(\alpha, \beta),$$

where α , M and κ are fixed, and β is a hyper-parameter:

$$\beta \sim \text{Gamma}(g, h).$$

Specifically, we take $\delta = 1$, $\alpha = 2$ (see Chap. 6 of Frühwirth-Schnatter, 2006 for a justification), $g = 0.2$, $h = 100g/\alpha R^2$, $M = \bar{y}$, and $\kappa = 4/R^2$, where \bar{y} and R are, respectively, the empirical mean and the range of the observed sample.

Regarding the application of free energy methods to univariate Gaussian mixture posterior distributions, Chopin et al. (2010) find that the two following functions of θ are efficient reaction coordinates: $\xi(\theta) = \beta$, and the potential function $V(\theta) = -\log \{p(\theta)p(y_{1:D}|\theta)\}$, that is, up to a constant, minus log the posterior density. However, the latter reaction coordinate is less convenient, because it is difficult to determine in advance the range $[x_{\min}, x_{\max}]$ of exploration. This is even more problematic in our sequential context. Using the IBIS strategy for instance, one would define $V_t(\theta) = -\log \{p(\theta)p(y_{1:t}|\theta)\}$, but the range of likely values for V_t would typically be very different between small and large values of t . Thus we discard this reaction coordinate.

In contrast, as discussed already in Chopin et al. (2010), it is reasonably easy to determine a range of likely values for the reaction coordinate $\xi(\theta) = \beta$. In our simulations, we take $[x_{\min}, x_{\max}] = [R^2/2000, R^2/20]$, where, again, R is the range of the data. Chopin et al. (2010) explains the good performance of this particular reaction coordinate as follows. Large values of β penalise small component variances, thus forcing β to large values leads to a conditional posterior distribution which favours overlapping components, which may switch more easily.

5.2.2. Numerical example

We consider the most challenging example discussed in Chopin et al. (2010), namely the Hidalgo stamps dataset, see e.g. Izenman and Sommer (1988) for details, and $K = 3$. In particular, Chopin et al. (2010) needed about 10^9 iterations of an Adaptive MCMC sampler (namely, an ABF sampler) to obtain a stable estimate of the free energy.

We run SMC samplers with the following settings: the number of particles is $N = 2 \times 10^4$, the criterion for triggering resample-move steps is $\text{ESS} < 0.8N$, and a move step consists of 10 successive Gaussian random walk steps, using the automatic calibration strategy described in Section 2.2.

We first run a SMC sampler, without free energy biasing, and using the IBIS strategy. Results are reported in Figures 1 and 2: the output is not symmetric with

respect to label permutation, and only one modal region of the posterior distribution is visited.

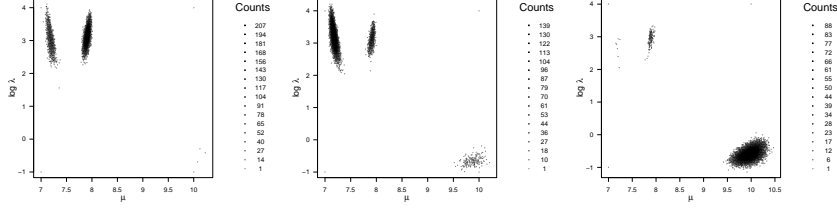


Figure 1: Hexagon binning for $(\mu_k, \log \lambda_k)$, $k = 1, 2, 3$, for the standard SMC sampler, no free energy biasing, IBIS strategy.

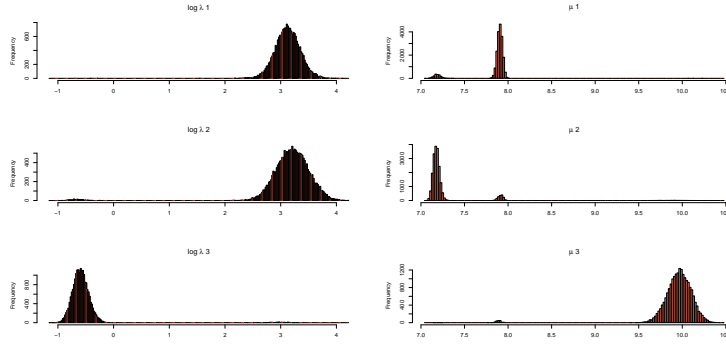


Figure 2: Weighted 1D histograms for the standard SMC sampler, no free energy biasing, IBIS strategy.

We then run a free energy SMC sampler, using the reaction coordinate ξ , 50 bins, and the ABP strategy for estimating the free energies. Figures 3 and 4 represent the cloud of particles before the final unbiasing step, when the particles target the free energy biased density $\tilde{\pi}_T$. Figures 5 and 6 represent the cloud of particles at the final step, when the target is the true posterior distribution. One sees that the output is not perfectly symmetric (at least after the final debiasing step), but at least the three equivalent modes have been recovered, and one can force equal proportions for the particles in each modal region, by simply randomly permuting the labels of each particles, if need be.

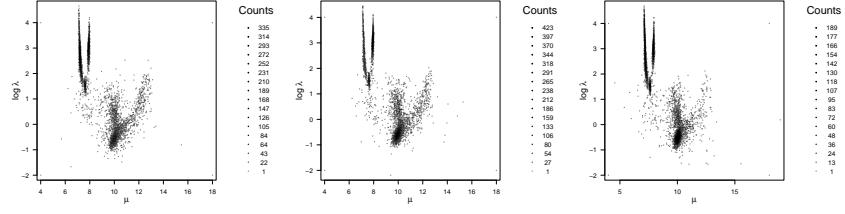


Figure 3: Hexagon binning for $(\mu_k, \log \lambda_k)$, $k = 1, 2, 3$, for the free energy SMC sampler, before the final debiasing step, IBIS strategy.

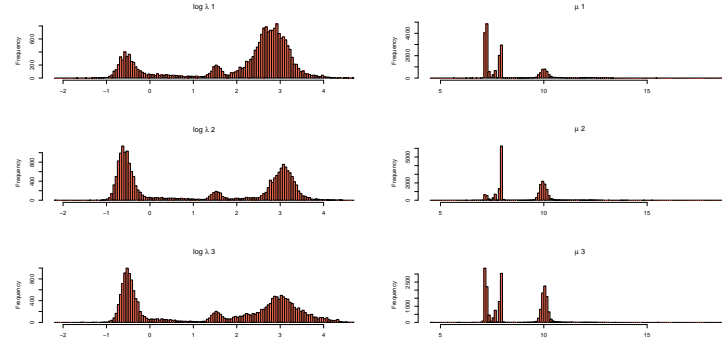


Figure 4: Histograms of the components of the simulated particles obtained by free energy SMC sampler, before the final debiasing step, IBIS strategy.

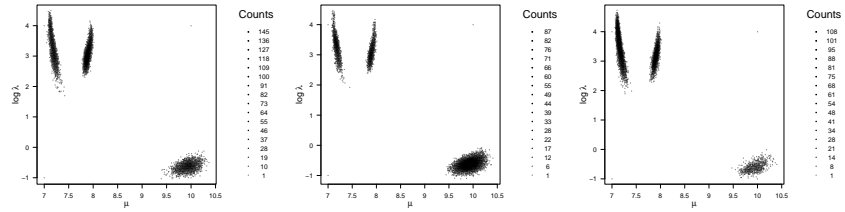


Figure 5: Hexagon binning for $(\mu_k, \log \lambda_k)$, $k = 1, 2, 3$, for the free energy SMC sampler, after the final debiasing step, IBIS strategy.

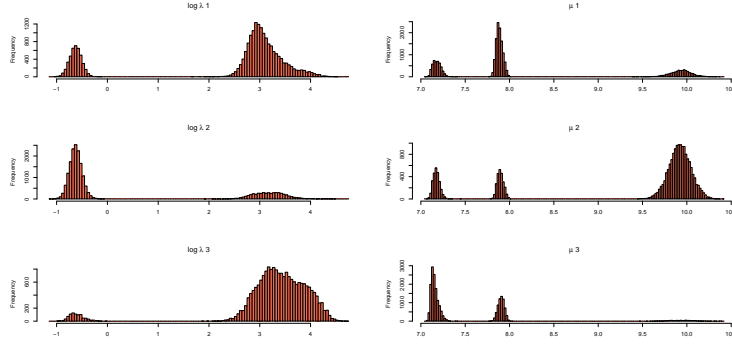


Figure 6: Histograms of the components of the simulated particles obtained by free energy SMC sampler, after the final debiasing step, IBIS strategy.

To assess the stability of our results, we run the same sampler ten times, and plot the ten so-obtained estimates of the overall free energy A_T , which is used in the last debiasing step; see Figure 7. Since a free energy function is defined only up to an additive function, we arbitrarily force the plotted functions to have the same minimum.

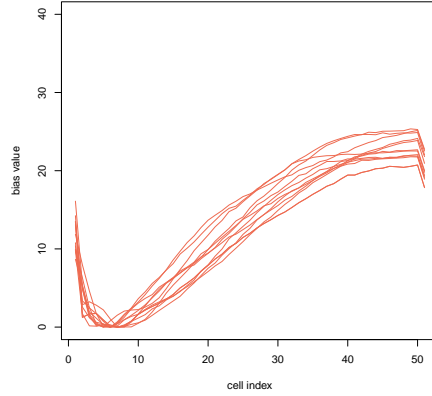


Figure 7: Estimates of the final free energy A_T obtained from 10 runs of a free energy SMC sampler, versus cell indices

In short, one sees in this challenging example that (a) a nearly symmetric output is obtained only if free energy biasing is implemented; and (b) using free energy SMC, satisfactory results are obtained at a smaller cost than the adaptive MCMC sampler used in Chopin et al. (2010).

5.3. Bivariate Gaussian mixtures

5.3.1. Prior, reaction coordinates

We now consider a bivariate Gaussian mixture, $\psi(y; \xi) = N_2(\mu, Q^{-1})$, which is parametrised as follows:

$$\xi_k = (\mu_{1,k}, \mu_{2,k}, d_{1,k}, d_{2,k}, e_k), \quad C_k = \begin{pmatrix} d_{1,k}^{1/2} & 0 \\ e_k & d_{2,k}^{1/2} \end{pmatrix}, \quad Q_k = C_k C_k^T.$$

This parametrisation is based on Bartlett decomposition: taking $d_{1,k} \sim \text{Gamma}(\alpha/2, \beta)$, $d_{2,k} \sim \text{Gamma}((\alpha - 1)/2, \beta)$, $e_k | \beta \sim N(0, 1/\beta)$ leads to a Wishart prior for Q_k , $Q_k \sim \text{Wishart}_2(\alpha, \beta I_2)$. This parametrisation is also convenient in terms of implementing the automatically tuned random walk Hastings-Metropolis strategy discussed in Section 2.2.

To complete the specification of the prior, we assume that

$$\mu_k = (\mu_{1,k}, \mu_{2,k})' \sim N_2(M, S^{-1}),$$

that $\alpha = 2$, and that $\beta \sim \text{Gamma}(g, h)$. Of course, this prior is meant to generalise the prior used in the previous section in a simple way. In particular, the hyper-parameter β should play the same role as in the univariate Gaussian case, and we use it as our reaction coordinate.

5.3.2. Numerical results

We consider two out of the four measurements recorded in Fisher's Iris dataset, petal length and petal width, see e.g. Frühwirth-Schnatter (2006, Chap. 6), and Figure 8 for a scatter-plot. We take $K = 2$. As in the previous example, we run a standard SMC sampler (with the same number of particles, and so on), and observes that only one mode is recovered. We then run a free energy SMC sampler. For the sake of space, we report only the debiased output at the very final stage of the free energy SMC sampler, that is the cloud of particles targeting the true posterior distribution. Figure 9 represents the bivariate vectors μ_k , and Figure 10 represent the component probabilities $q_k = \omega_k / (\omega_1 + \omega_2)$ for $k = 1, 2$. Clearly, the output is nearly symmetric.

One sees in this example that free energy SMC still works well for bivariate Gaussian mixture model, despite the larger dimension of the parameter space. In particular, the choice of the reaction coordinate seems to work along the same lines, i.e. choosing an hyper-parameter that determines the spread of the components.

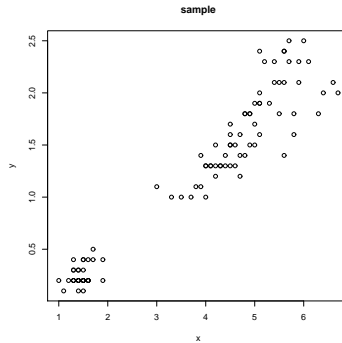


Figure 8: Iris sample

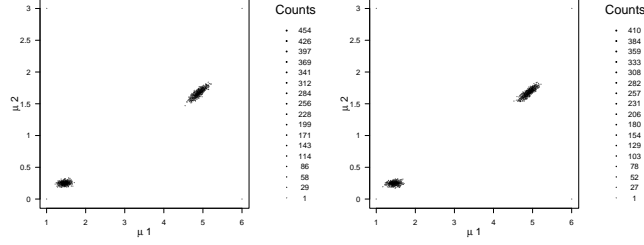


Figure 9: Hexagon binning for $\mu_k = (\mu_{k,1}, \mu_{k,2})$, $k = 1$, bivariate Gaussian example

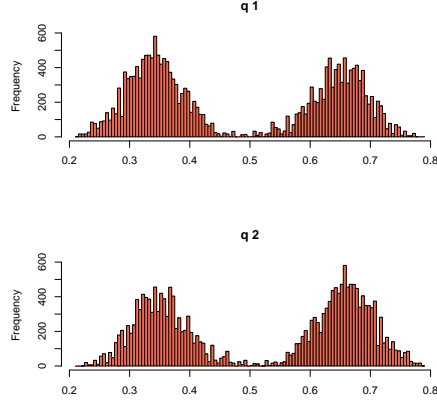


Figure 10: Weighted histograms of $q_k = \omega_k / (\omega_1 + \omega_2)$, for $k = 1, 2$, bivariate Gaussian example

6. CONCLUSION

In this paper, we introduced free energy SMC sampling, and observed in one mixture example that it may be faster than free energy methods based on adaptive MCMC, such as those considered in Chopin et al. (2010). It would be far-fetched to reach general conclusions from this preliminary study regarding the respective merits of free energy SMC versus free energy MCMC, or, worse, SMC versus Adaptive MCMC. If anything, the good results obtained in our examples validates, in the mixture context, the idea of combining two recipes to overcome multimodality, namely (a) free energy biasing, and (b) tracking through SMC some sequence (π_t) of increasing difficulty, which terminates at $\pi_T = \pi$. Whether such combination should work or would be meaningful in other contexts is left for further research.

ACKNOWLEDGEMENTS

N. Chopin is supported by the 2007–2010 grant ANR-07-BLAN “SP Bayes”. P. Jacob is supported by a PhD Fellowship from the AXA Research Fund. The authors thank Peter Green for insightful comments.

REFERENCES

- Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statist. Comput.*, 18(4):343–373.
- Carpenter, J., Clifford, P., and Fearnhead, P. (1999). Improved particle filter for nonlinear problems. *IEE Proc. Radar, Sonar Navigation*, 146(1):2–7.
- Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Am. Statist. Assoc.*, 95:957–970.
- Chopin, N. (2002). A sequential particle filter for static models. *Biometrika*, 89:539–552.
- Chopin, N., Lelièvre, T., and Stoltz, G. (2010). Free energy methods for efficient exploration of mixture posterior densities. *Arxiv preprint arXiv:1003.0428*.
- Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436.
- Diebolt, J. and Robert, C. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. R. Statist. Soc. B*, pages 363–375.
- Doucet, A., de Freitas, N., and Gordon, N. J. (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York.
- Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *J. Am. Statist. Assoc.*, 96(453):194–209.
- Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer.
- Gelman, A. and Meng, X. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185.
- Gilks, W. R. and Berzuini, C. (2001). Following a moving target - Monte Carlo inference for dynamic Bayesian models. *J. R. Statist. Soc. B*, 63:127–146.
- Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F, Comm., Radar, Signal Proc.*, 140(2):107–113.
- Izenman, A. J. and Sommer, C. J. (1988). Philatelic mixtures and multimodal densities. *J. Am. Statist. Assoc.*, (83):941–953.
- Jasra, A., Holmes, C., and Stephens, D. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statist. Science*, pages 50–67.
- Jasra, A., Stephens, D., and Holmes, C. (2007). On population-based simulation for static inference. *Statistics and Computing*, 17(3):263–279.
- Kong, A., Liu, J. S., and Wong, W. H. (1994). Sequential imputation and bayesian missing data problems. *J. Am. Statist. Assoc.*, 89:278–288.
- Lelièvre, T., Rousset, M., and Stoltz, G. (2010). *Free-energy computations: a mathematical perspective*. Imperial College Press.
- Liu, J. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *J. Am. Statist. Assoc.*, 93:1032–1044.
- Marin, J. and Robert, C. (2007). *Bayesian core: a practical approach to computational Bayesian statistics*. Springer Verlag.

- Neal, R. M. (2001). Annealed importance sampling. *Statist. Comput.*, 11:125–139.
- Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. R. Statist. Soc. B*, 59(4):731–792.
- Whitley (1994). A genetic algorithm tutorial. *Statist. Comput.*, 4:65–85.